LEARNING TO SEE: RESEARCH IN TRAINING A ROBOT VISION SYSTEM

Gary Witus*
Turing Associates, Ann Arbor, MI 48103

Robert E. Karlsen
U.S. Army –TARDEC, Warren, MI 48397-5000

ABSTRACT

Engineered or hard-coded autonomous behaviors tend to be "brittle," working for a narrow range of conditions but failing outside that range. Trainable robots capable of learning and adapting to new environments and conditions have the potential for greater robustness and reusability. Trainable robots would not be restricted to learning from their own experience, but could potentially integrate models or lessons learned by other similar robots operating in different conditions, thus achieving a "learning force multiplier". In this research we began an investigation of issues and methods in robot learning, from definition of the learning objective, training methods, learning algorithms, and integration of models or lessons from multiple training sessions. Our objective in this initial research was not to develop new robot learning technologies, but to explore issues and approaches across all aspects of robot learning. In this stage of the project, we focused on learning to see, specifically learning to discriminate between "Go" and "NoGo" terrain.

1. INTRODUCTION

At the present time, the vast majority, if not all, of the mobile ground robots in use by the military in the field are teleoperated. Despite widespread research and development in university and government laboratories, autonomous and semi-autonomous robots have yet to gain field acceptance. This is due in large part to concern that the systems would fail to perform correctly in the highly varied and unpredictable field environment, resulting in possible mission failure and/or safety risk.

Engineered solutions to autonomous driving behaviors tend to be "brittle." They may work in narrowly defined set of conditions, e.g., an office, a laboratory, or even lunar or Martian terrain, but fail and not recover when placed in the various terrestrial conditions.

Trainable robots, i.e., robots capable of learning the characteristics of new environments and adapting their behavior to accommodate those characteristics, would be an important step towards the development of robust and effective semi-autonomous systems. A trainable robotic system can potentially learn not only from its own experience, but can assimilate the lessons from other similar systems, thus greatly increasing the domain of operation.

The objective of this research was to explore the spectrum of issues and technologies in a robot vision system capable of being trained to recognize the trafficability of widely varied terrain types and features, in unstructured outdoor conditions.

Learning for visual terrain recognition is an emerging area of robotics research. Sebastian Thrun (2005, 1996), lead researcher on the Stanford team that won the DARPA Grand Challenge, has been a long time advocate of robot learning to produce robust, adaptable, and transferable robot behaviors. Research in this area has been limited, and has not addressed robust, natural world conditions. Work by Howard and Seraji (2001) and Howard et al. (2001) was focused on barren extraterrestrial terrain conditions, with the complexities of vegetation, man-made structures, and water. Earlier work by Karlsen and Witus (2008; 2007; 2007) addressed terrestrial learning for terrain classification using local pixel properties, not large-scale structure, and supervised learning with either a priori terrain classification or trainer-assigned trafficability. Angelova et al. (2007) describe a slip-prediction system, but which was tested against a set of distinctive a priori terrain types.

2. APPROACH

Our approach involved collecting training data in a robust and varied environment for subsequent investigation of component learning methods.

The first step was to define an approach to collect training data. We decided that we wanted the robot to learn by observing human control behavior, i.e., from the actions that a human operator takes during exercises. An unsupervised learning system will have vastly more opportunities to learn and more relevant experience than a supervised learning system that requires human intervention.

We realized that the training data had to be collected in a manner similar to the intended behavior of the robot

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01 DEC 2008	**N/A**	**-**

4. TITLE AND SUBTITLE	
Learning to See: Research in Training a Robot Vision System	5a. CONTRACT NUMBER **W56HZV-06-C-0577**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) **Gary Witus; Robert E. Karlsen**	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Turning Associates, Ann Arbor, MI 48103 US Army RDECOM-TARDEC 6501 E 11 Mile Rd Warren, MI 48397-5000**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) **US Army RDECOM-TARDEC 6501 E 11 Mile Rd Warren, MI 48397-5000**	10. SPONSOR/MONITOR'S ACRONYM(S) **TACOM/TARDEC**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S) **19371**

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
Presented at the 26th Army Science Conference, 1-4 December 2008, Orlando, Florida, USA, The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**SAR**	**8**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

system. In our concept, the robot would be given a sequence of waypoints, and be expected to drive from waypoint to waypoint as directly as possible while avoiding obstacles and rough terrain. This defined our data collection approach.

The second step was to select a data collection platform. For this purpose we instrumented an off-road vehicle (specifically, a tractor) with a differential GPS, 6 degree of freedom inertial measurement unit (IMU), and a monochrome stereo camera system. Data from the sensors were logged on an on-board computer. The IMU data were logged at 20 Hz, the stereo camera frames were logged at 7.5 Hz, and the GPS data were logged at 1 Hz. The GPS unit has a long-run accuracy of one meter, as determined from a stationary antenna over a 24-hour period as satellites crossed the horizon. It has an RMS sample-to-sample variation of approximately 10 cm. The stereo camera system had a 60 degree by 40 degree field of view, 640 by 480 pixel resolution, and was mounted at 63 inches elevation, pointing slightly down so that the angle from vertical to bottom row of the image was 57 degrees. On a flat surface, a point at the bottom of the image would be 53 inches in front of the camera.

In order to have value for training, the segments from waypoint to waypoint had to require turning to avoid obstacles and rough terrain. We selected a sequence of waypoints from a 1-foot resolution aerial of the test site. The waypoints were selected to present a variety of terrain features and characteristics including fields, hills, ditches, ponds, streams, woods, isolated bushes and trees, roads, fallen logs, fences, trucks, and buildings.

The third step was to collect the data. The data were collected in conditions of full summer daylight with partial, intermittent cloud cover. We collected data on three runs on different days, as shown in figure 1 (the aerial photograph had a slight warp relative to the GPS track of the data collection). Each run lasted approximately 45 minutes, and collected 10 GB of data. At the present time, we have only analyzed the data collected in the run shown in figure 2. The waypoint segments in figure 2 are shown in alternating red and yellow. In some cases, the waypoint segments overlap, in reverse direction.

Fig. 1: Test Area And Routes

Fig. 2: Route 1 Waypoint Segments

The fourth step was to reduce the data for analysis. We reduced the data to 1 Hz. For each one second interval we computed the mean and standard deviation of the IMU outputs, and selected stereo pairs at one second intervals. The IMU outputs were the roll, pitch and yaw angles and angular rates, and the X, Y and Z accelerations and rates. The GPS data were the latitude, longitude and elevation (at 0.25 cm precision, but only 10 cm accuracy). For each time step we computed the bearing from the current position to the upcoming waypoint.

We computed a disparity image from each stereo image pair. At this point we had to decide what and how much engineered or hard-coded visual processing to perform prior to sending the input to the trainable soft-computing engine, i.e., to select the representation of the visual input. We did not want to prejudge what features were of interest; this is what we wanted the soft computing system to discover. However, without some pre-processing the input would be too varied and the learning system would be confronted with too deep of a learning problem for the size of the data set.

We considered several biologically-inspired options: simple retino-topic disparity and luminance images, multi-resolution retino-topic disparity and luminance images, and multi-resolution oriented receptive field disparity and luminance images. The engineering alternative was to compute a point cloud, i.e., elevation as a function of distance and heading relative to the forward direction. For the initial investigation, we decided to use the point cloud perception of the terrain as the input to the learning system.

The significant roll of the terrain presented a challenge computing a useful elevation map (see figure 3). At issue is the ground plane. From a driving perspective, the local variation in elevation (vertical texture) and the elevation of features relative to the local ground plane at the feature are important. Elevation relative to the ground plane of the vehicle at its location when it captures the image is irrelevant.

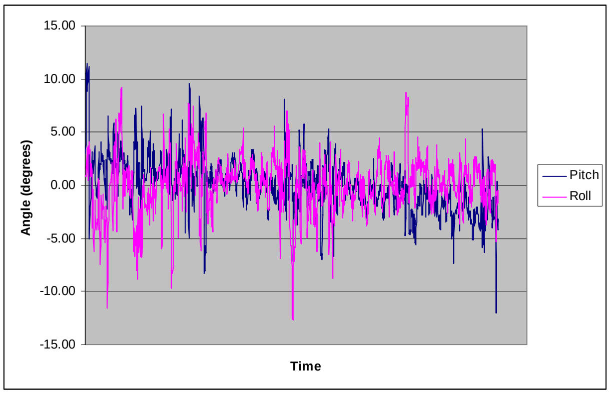

Fig. 3: Pitch And Roll

To resolve this issue, we computed the standard deviation of the elevation with distance and angle bins from the point cloud. This is a measure of the variation in elevation over a grid cell in polar coordinates. Using this representation required us to define the resolution of the representation, i.e., the size of the bins, and opens the door to multi-resolution representations. We computed the elevation standard deviation maps at several coarse resolutions (3-by-3, 9-by-9, 12-by-12, and 15-by-15). This representation constituted feature vectors for training. The extents of the maps were distances from 4 to 24 feet from the vehicle, and from plus-30 to minus-30 degrees from the current heading. We also computed the luminance and luminance variation as functions of distance and heading, but did not use luminance data in this stage of the modeling and analysis.

We inferred trafficability of the upcoming terrain from the driver's behavior relative to the upcoming waypoint. We reasoned that if the vehicle was being driven more or less forward, the upcoming terrain must be trafficable. If the bearing to the upcoming waypoint were more or less straight ahead, and the vehicle were turning to the right or left, then the upcoming terrain must be un-trafficable. If the vehicle were turning to the right and the upcoming waypoint were on the left (and vice-versa), then the upcoming terrain must be un-trafficable. The classification matrix is shown in figure 4 ("X" denotes cases in which we could not infer trafficability of the upcoming terrain).

Heading Relative To Waypoint	Steering Action:	Left	Center	Right
	Left		Go	NoGo
	Center	NoGo	Go	NoGo
	Right	NoGo	Go	

Fig. 4: Trafficability Classification Matrix

On this basis, we constructed a fuzzy logic model to compute the trafficability of the upcoming terrain from the driver behavior (see figure 5). We set the deadband around zero yaw rate, defining "not turning," as being one standard deviation of the yaw rate from several segments where vehicle was intended to be on a straight path, and the threshold for "definitely turning" at three times the standard deviation. We set the deadband around zero for the waypoint to be considered straight ahead to be 10 degrees or one sixth the width of the image, and the threshold for the waypoint to be definitely not straight ahead to be three times that.

The fifth step was to train an algorithm to predict the trafficability score (the output of the fuzzy logic model) from the local perceived terrain feature vector. We began with a low-resolution 3-b-3 grid, producing a length 9 feature vector. We used a simple multi-linear regression model to establish a baseline prediction capability. We have not yet completed analysis using decision tree and artificial neural network methods. We have not yet examined higher resolution feature vectors, or feature vectors combining luminance and point cloud data.

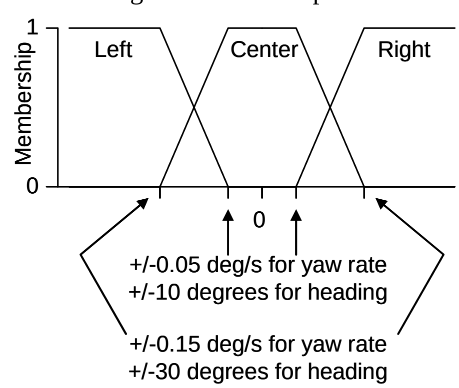

Fig. 5: Fuzzy Logic Models For Steering Action and Relative Waypoint Heading

The sixth and final step was to attempt to combine models from different segments. The basic approach was to model the appearance of the training data for each segment. Given some new observation, these models would be used to evaluate how much the observation resembles each of the training data sets, and thus weight the predictions of the various models. We are still investigating formulations to distinguish the training data using statistical, decision tree, and artificial neural

network methods.

3. RESULTS

3.1 Data Collection

We found that it was virtually impossible to drive in rough off road conditions restricted only to visual input in the frontal 60 by 40 degrees. Despite best intentions, we sometimes needed to look to the side and steeply down. We also found it difficult to drive without remembering what objects had been to the right or left, but which were now out of the frontal field of view. We explored accommodating this by putting a lag of the perception vector to synchronize it with the driving behavior, but did not find any consistent lead-lag relationship. We also observed that, on the rough terrain, ground steering was significant, and the driver often over-corrected. These effects and behaviors were not functions of the forward visual perception, and added some amount of noise to the input to the fuzzy logic trafficability assessment model.

3.2 Data Reduction

The GPS/IMU data reduction was relatively straightforward and uneventful. We encountered multiple problems with the stereo camera data. One problem was limited dynamic range. The summer day scenes contained very bright illumination and dark shadows. Paths through the woods were entirely in shadow, others were mixed sun and shadow. The camera's automatic gain control attempted to compensate for this, but with limited dynamic range, portions of the images were often overexposed (white) and underexposed (black). The automatic gain control for the right and left cameras were not identical. Thus different portions of the right and left images would be overexposed and underexposed. Naturally this would degrade stereo disparity calculation and lead to erroneous matches. Several examples of the luminance images are shown in figure 6.

Fig. 6: Example Scene Images

We used a large kernel to compute the stereo disparity (40 pixels). For the most part this produced good results and a dense, reduced resolution disparity image. However, when the stereo images had significantly different over and under exposures, or when the images had large over and under exposed regions, the stereo disparity calculations produced garbage output. We did not come up with a robust and reliable approach to detect the problem conditions. Instead we applied an adaptive median filter to smooth over disparity calculation errors. This had the unavoidable side effect of blurring some shape details. Figure 7 shows example disparity images corresponding to the luminance images in figure 6.

3.3 Trafficability Inference

The model to infer trafficability from steering action (from the yaw rate, to be precise) and difference between current heading and bearing to the upcoming waypoint was largely consistent with after-the-fact human judgment. We stepped through the imagery at 60 second intervals and displayed the computed trafficability. In some cases, it was not possible to visually determine the direction to the waypoint accurately (resulting in errors during driving). In some cases the steering actions were due to ground steering, or in response to ground steering, or due to steep pitch or roll angles, and not in response to upcoming terrain perception. But the computed trafficability index was in general agreement with the subjective check.

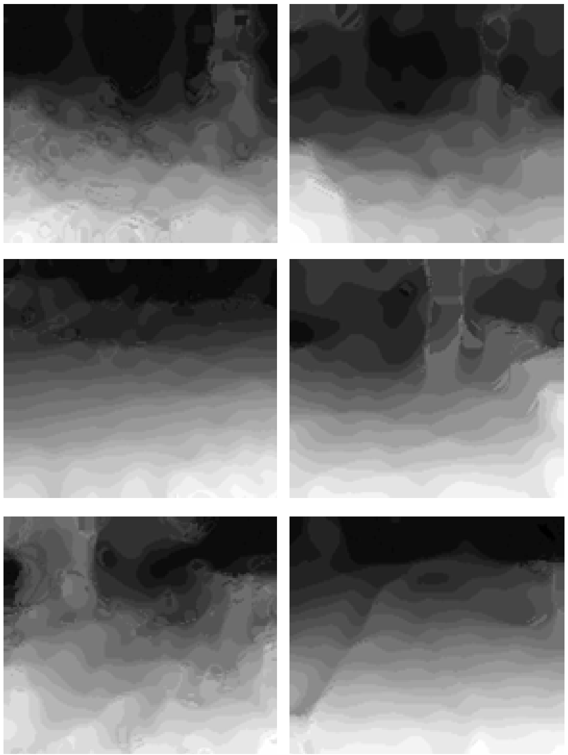

Fig. 7: Example Disparity Images

3.4 Modeling Trafficability as a Function of Perception

We employed a two-stage approach to model trafficability inferred from driving action as a function of the stereo perception feature vector. The first stage was local learning applied to individual waypoint segments to discover high-level appearance features that were useful to characterize trafficability, and their associated trafficability expectation. The second stage was sequential learning which combined the high-level features and trafficability expectations from the individual waypoint segments.

In the local learning first stage, we assumed that there would be periods of time during which the terrain appearance and trafficability would not vary significantly, causing the data to consist of sub-segments with relatively similar appearance and trafficability. We cut the segment into sub-segments whenever either (a) the Euclidean distance between the sequential feature vectors exceeded a threshold, or (b) the distance between the inferred trafficability exceeded a threshold. For each sub-segment, we computed the mean and standard deviation of the appearance vector and of the trafficability score. We then consolidated non-adjacent sub-segments if both the Euclidean distance between the mean feature vectors and the distance between the mean trafficability scores were below their respective thresholds, and iterated the consolidation until all consolidation was completed. The mean appearance feature vectors and their associated trafficability expectation constituted the first pass at identifying the high-level features.

Once we identified the high-level features and their associated trafficability expectation for a given waypoint segment, we evaluated the predictability of the trafficability score from the appearance vector. The purpose of this step was to identify those segments that were inherently unpredictable, so that they could be excluded from further use in the analysis. For some segments, the variation in trafficability score was due to factors not represented in the terrain appearance vector. In some cases when the terrain had high slope, steering was to get to more level terrain. In some cases, the driver could not see the location of the next waypoint, which caused erratic steering. In some cases, the driver stayed on the gravel road rather than cut across the lawn at the corner, not because it was too rough but because it was someone’s lawn. In some cases the steering action was in response to an object that was outside the field of view of the camera, too close and/or to the side. In some cases, vehicle yaw was due to ground-steering effects. All of these cases produced an inferred trafficability score that reflected factors other than the distribution of elevation of the upcoming terrain, and therefore should be excluded from use in the modeling and analysis.

For a given observation, i.e., an appearance vector, we forecast the trafficability score as the weighted average of the mean trafficability score for each sub-segment. The weighting factor was one divided by the Euclidean distance between the observation appearance vector and the mean appearance vector for the sub-segment (plus a small epsilon to avoid the possibility of dividing by zero). We eliminated from further use those segments in which the residual error was greater than the trafficability score threshold used in separating sub-segments and consolidating the high-level feature vectors.

We did not have a prior expectation of the appropriate threshold values, and therefore conducted a parametric analysis varying the levels of the two thresholds. Over all non-rejected segments and high-level features, we computed the proportion of variance in trafficability score accounted for by the high-level features (i.e., r-squared), the number of segments rejected, and the representation ratio (the number of high-level features relative to the number of segments accepted). Figures 8, 9 and 10 show the results of predicting the trafficability score from the high-level features associated with each segment. Figure 8 shows the fraction of waypoint segments not rejected. Figure 9 shows the average number of high-level features per waypoint segment. Figure 10 shows the proportion of variance in trafficability score over all the non-rejected segments explained (r-squared) by the pool of high-level features.

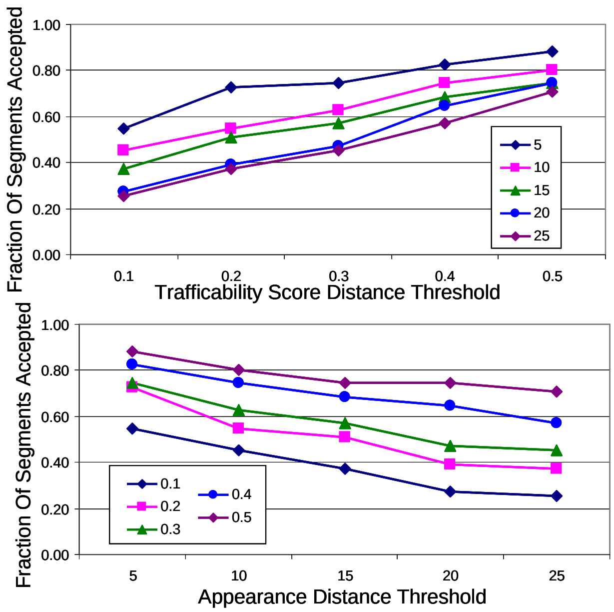

Fig. 8: Proportion of Waypoint Segments Not Rejected

Requiring tight trafficability score tolerance increased the explanatory power (r-squared), decreased the number of features per segment, and decreased the fraction of segments rejected. Allowing broadly defined feature appearance, reduced the explanatory power, decreased the feature ratio, and increased the fraction of segments rejected.

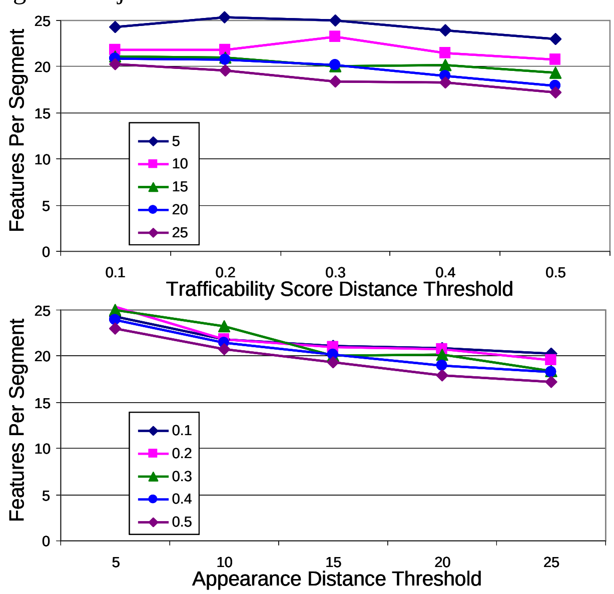

Fig. 9: Mean Number of High-level Features Per Segment Before Cross-Segment Consolidation

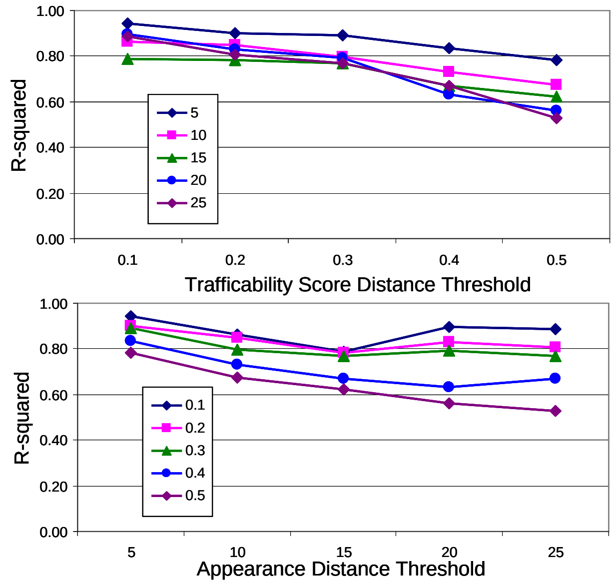

Fig. 10: Trafficability Score R-squared Before Cross-Segment Consolidation

We next consolidated the high-level features across segments, using the same threshold values used to cut the segments into sub-segments, and to consolidated high-level features within each segment. The performance of the consolidated model is presented in figures 11 and 12. Figure 11 shows the mean number of high-level features per non-rejected waypoint segment. Figure 12 shows the proportion of variance in trafficability score over all the non-rejected segments explained (r-squared) by the pool of high-level features.

Consolidation across segments had more dramatic effects on the proportion of variance explained by the model (r-squared) than on the number of high-level features per segment. The effects of cross-segment consolidation were sensitive to the trafficability score distance threshold, but insensitive to the appearance distance threshold.

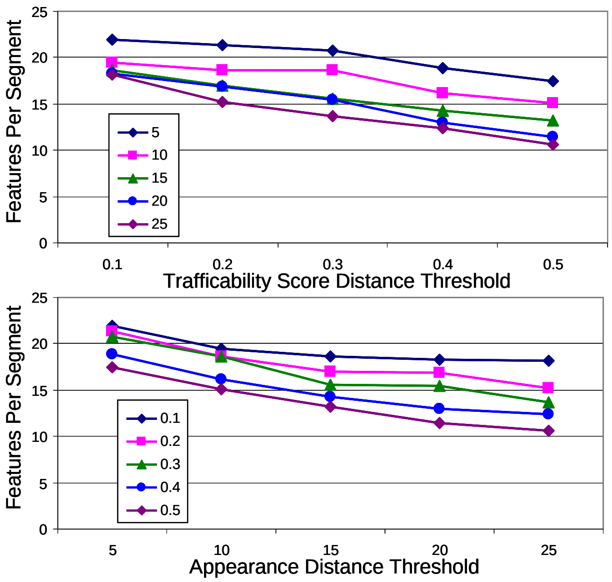

Fig.11: Mean Number of High-level Features Per Segment After Cross-Segment Consolidation

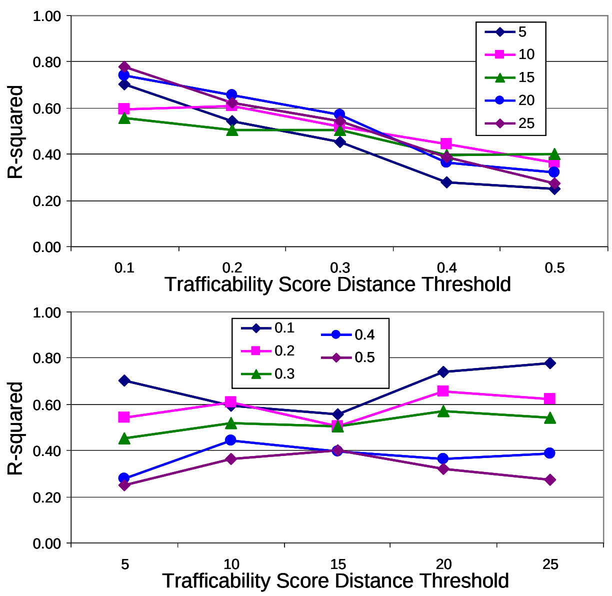

Fig. 12: Trafficability R-squared After Cross-Segment Consolidation

These results show that by setting the thresholds used in defining the high-level features, we can automatically extract high-level features for the set of route segments that are strongly correlated with trafficability score, provided we restrict to route segments that are self-explanatory (i.e., internal to the route segment, trafficability score and the appearance vector are correlated). When we choose a large distance threshold for separating the appearance of high level features and a small threshold for separating trafficability score, the trafficability score associated with the closest high level feature explains 78 percent of the variance in trafficability score over the self-explaining route segments. The number of features per segment (18) is in the middle range. The low threshold on trafficability distance tends to increase the number of features, but the large threshold on difference in appearance tends to reduce the number of features. However, only 25 percent of the route segments met the conditions of being self-explanatory for these threshold settings.

CONCLUSIONS

This paper reports on work in progress in the emerging and challenging area of developing trainable robots, and methods for training robots. While we have not achieved significant technical success, we have learned several significant observations and lessons for future research.

For a robot to be able to learn from operator behavior in training situations, the basis for the operator's decisions and the information available to the robot behavior need to be consistent. If the operator is using wide field of view input (or head movement), that same field of view must be available to the robot. If the operator is basing decisions on short term memory of the local surrounds, comparable short term memory representations must be built into the robot system. If the operator sometimes bases driving decisions on the slope of the current terrain, or the angle with respect to the sun, then these data should be made available to the learning system. If there are factors that influence the driver's decisions that are not sensory inputs to the learning system, then the ability of the learning system to predict driver behavior will be limited and the possibility of learning spurious associations will be increased.

For a robot system to learn by imitation, the human operator must exhibit the desired behaviors during training runs. If the human operator can not exhibit the behaviors desired of the robot, the conceptual robot behaviors should be re-evaluated.

In operations in natural illumination on rough terrain, there will be times when the sensor input is degraded, e.g., due to motion blur, solar glare, etc. The robot systems need to be able to detect degraded input, and proceed based on prior assessments. An important part of a robust robot system will be adaptive behaviors for conditions of degraded perception.

The learning and training system needs to be able to infer terrain trafficability or hazardousness from the operator behavior. We can not afford to have the operator drive off a cliff or into a fire in order to demonstrate that these are undesirable actions.

The principles of simplicity and consistency apply to robot training just as they do for animal or human

training. If individual segments involve a wide variety of terrain types and obstacle, the learning systems will have difficulty with the data. Complex, free play lessons are undesirable. An ideal would be for a training lesson to have one type of good terrain and one type of obstacle or rough terrain: advance, and then sharply turn away from the obstacle or rough terrain. The turn away should be at a consistent distance.

General-purpose machine learning methods, such as artificial neural networks, decision trees, and multi-linear regression models do not have an inherent representation of 2-D or 3-D spatial relationships, or other structure to the input. Each dimension of the feature vector is independent, as far as the representation is concerned. If the feature vector actually represents a grid or retino-topic map or multi-resolution map, there is no way to represent this to the learning mechanism. The patterns that we humans see in the input are in the context of the structure of the input representation. General-purpose machine learning systems are at a relative disadvantage: they have to infer a pattern without knowing the underlying grammar or geometry. Converting a two dimensional image into a one dimensional vector before inputting it to a learning algorithm strips out important structuring information. It is like stripping out all the spaces and punctuation marks from text before trying to learn a foreign language. Research in machine learning methods with inherent geometrical structures is needed.

Our approach to learning to predict trafficability from appearance involved discriminating between those route segments that were self-explanatory from those that were inherently unpredictable. For self-explanatory segments, when the segment was analyzed in isolation, clustering based on the appearance vector explained the variance in trafficability. For inherently unpredictable segments it did not. An important part of a practical robot intelligence system is the ability to discern when its predictions or decisions are reliable and when they are not, i.e., when its algorithms and data are applicable and when they are not. A system that does not make this discrimination is at risk of making inappropriate decisions without knowing it.

Our research focused on automatically extracting high-level features associated with driving behaviors. The high-level features were defined in the framework of low-level feature vector. In this research we used a 9 dimensional feature vector, the standard deviation of elevation within a distance-by-azimuth grid cell. The high-level features are contingent on the low-level feature vector. Further research is needed into alternative or complementary low-level feature vectors for visual perception. Other factors such as color or luminance variation could be incorporated. Resolution effects need to be studied. Retino-topic representations (i.e., in the original image plane) should be investigated. Incorporating complementary non-visual sensory data, e.g., pitch and roll angles, should also be explored. The value and potential contribution of a learning system is for the system to automatically determine how to use and combine the disparate feature vector data.

Our approach relied on multi-stage clustering or learning methods. It exploited the sequential nature of the input stream and the assumption of local self-similarity of the terrain. It performed fast clustering using sub-segmentation followed by consolidation with a route segment. On this basis, it could determine whether or not the segment met the self-explanatory requirement. For those segments that met the requirement, their high-level features were consolidated with the high-level features extracted from prior self-explanatory route segments. This produced fast assimilation of new lessons.

REFERENCES

Angelova, A., L. Matthies, D. Helmick, P. Perona. “Learning and prediction of slip from visual information.” Journal of Field Robotics 24(3), 205–231 (2007).

Howard, A., and H. Seraji, “Vision-based terrain characterization and traversability assessment,” *J. Robotic Systems* 18(10), 577-587 (2001).

Howard, A., E. Tunstel, D. Edwards and A. Carlson, “Enhancing fuzzy robot navigation systems by mimicking human visual perception of natural terrain traversability,” Joint 9th IFSA World Congress and 20th NAFIPS Int. Conf., Vancouver, B.C., Canada, 7-12, July 2001.

Karlsen, R., and G. Witus. “Adaptive learning applied to terrain recognition.” Proc. SPIE, Volume 6962, Intelligent and Autonomous Behaviors II. (2008); DOI:10.1117/12.777289.

Karlsen, R., and G. Witus. “Terrain Perception for robot navigation.” Proc. SPIE, Volume 6561, Perception. (2007); DOI:10.1117/12.720829.

Karlsen, R.E. and G. Witus, “Vision-based terrain learning,” *Unmanned Ground Vehicle Technology VIII*, SPIE Proc. 6230, 33-42 (2006).

Thrun, S., W. Burgard, and D. Fox. *Probabilistic Robotics*. MIT Press, Cambridge, MA, 2005.

Thrun, S. Explanation-Based Neural Network Learning: A Lifelong Learning Approach. Kluwer Academic Publishers, Boston, MA, 1996.